NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

AN ANALYSIS OF VESSEL WAYPOINT BEHAVIOR THROUGH DATA CLUSTERING

by

John R. Hintze

September 2017

Thesis Advisor:	Lyn R. Whitaker
Second Reader:	Robert A. Koyak

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			*Form Approved OMB No. 0704-0188*
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington DC 20503.			
1. AGENCY USE ONLY ***(Leave blank)***	**2. REPORT DATE** September 2017	**3. REPORT TYPE AND DATES COVERED** Master's thesis	
4. TITLE AND SUBTITLE AN ANALYSIS OF VESSEL WAYPOINT BEHAVIOR THROUGH DATA CLUSTERING			**5. FUNDING NUMBERS**
6. AUTHOR(S) John R. Hintze			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000			**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A			**10. SPONSORING / MONITORING AGENCY REPORT NUMBER**
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB number ____N/A____.			
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.			**12b. DISTRIBUTION CODE**

13. ABSTRACT (maximum 200 words)

In this thesis, we cluster stop points into stop-point regions using one month's Automatic Identification System (AIS) data from the Gulf of Mexico and Caribbean Sea to characterize vessel behavior in an area with diverse traffic patterns. Initial cleaning of the dataset is necessary to address multiple issues common to AIS transponders. We consider methods for computing inter-point distances. In particular, we study a promising method for combining geospatial coordinates with other vessel attributes. We use the Ordering Points To Identify the Cluster Structure (OPTICS) clustering algorithm because it can identify outliers, and it constructs clusters of varying shapes and densities. Our best results come from dividing the area of interest into seven zones of equal size, and analyzing the results over each zone. Using classification trees to develop a classification tool, we illustrate an approach for predicting the cluster membership of a new observation. Due to the reduction in computation time and accuracy of results, we recommend that further research utilize the methods from this study as the foundation for an automated threat detection system.

14. SUBJECT TERMS data analysis, automated identification system, clustering, anomaly detection			**15. NUMBER OF PAGES** 67
			16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	**18. SECURITY CLASSIFICATION OF THIS PAGE** Unclassified	**19. SECURITY CLASSIFICATION OF ABSTRACT** Unclassified	**20. LIMITATION OF ABSTRACT** UU

NSN 7540-01-280-5500 Standard Form 298 (Rev. 2-89) Prescribed by ANSI Std. 239-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

AN ANALYSIS OF VESSEL WAYPOINT BEHAVIOR THROUGH DATA CLUSTERING

John R. Hintze
Ensign, United States Navy
B.S., United States Naval Academy, 2016

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN APPLIED SCIENCE (OPERATIONS RESEARCH)

from the

NAVAL POSTGRADUATE SCHOOL
September 2017

Approved by: Lyn R. Whitaker
Thesis Advisor

Robert A. Koyak
Second Reader

Patricia Jacobs
Chair, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

In this thesis, we cluster stop points into stop-point regions using one month's Automatic Identification System (AIS) data from the Gulf of Mexico and Caribbean Sea to characterize vessel behavior in an area with diverse traffic patterns. Initial cleaning of the dataset is necessary to address multiple issues common to AIS transponders. We consider methods for computing inter-point distances. In particular, we study a promising method for combining geospatial coordinates with other vessel attributes. We use the Ordering Points to Identify the Cluster Structure (OPTICS) clustering algorithm because it can identify outliers, and it constructs clusters of varying shapes and densities. Our best results come from dividing the area of interest into seven zones of equal size, and analyzing the results over each zone. Using classification trees to develop a classification tool, we illustrate an approach for predicting the cluster membership of a new observation. Due to the reduction in computation time and accuracy of results, we recommend that further research utilize the methods from this study as the foundation for an automated threat detection system.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF FIGURES

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF ACRONYMS AND ABBREVIATIONS

AIS	Automatic Identification System
AOR	area of responsibility
CMRE	Centre for Maritime Research and Experimentation
CSV	comma separated values
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DHS	Department of Homeland Security
ETA	estimated time of arrival
IMO	International Maritime Organization
JSON	JavaScript object notation
MID	maritime identification digits
MMSI	maritime mobile service identity
NM	nautical miles
NMEA	National Maritime Electronics Association
NRC	National Research Council
OEL	overall error level
OPTICS	Ordering Points to Identify the Cluster Structure
S-AIS	Satellite Automatic Identification System
SOTDMA	Self-Organized Time Division Multiple Access
TREAD	Traffic Route Extraction and Anomaly Detection
UTM	Universal Transverse Mercator
VHF	very high frequency
VTS	Vehicle Traffic Service

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

Although originally conceived as a collision avoidance system, Automatic Identification Systems (AIS) have completely altered the way analysts look at maritime vessel data. Since the mandate in 2000 by the International Maritime Organization (IMO) for vessels meeting certain size and passenger requirements to be outfitted with AIS, organizations have been established to collect and maintain databases of raw AIS data (IMO 2017). Most topics studied in this field involve anomaly detection or path prediction using "ship tracks." Furthermore, such studies are generally performed on areas that constrain ship routes to very few ship lanes travelling to limited locations. We choose our area of responsibility as the Gulf of Mexico and Caribbean Sea, an area with high traffic density and populous ports located throughout the area. Analyzing the stop points over an area with difficult-to-classify vessel routes is a new approach to anomaly detection that could provide the foundation for an automated threat detection system that can be adapted for worldwide use.

We convert the dataset used in this study from its raw format into multiple comma separated value files. These files contain worldwide static and dynamic data from January 1−31, 2014. We define static data as the generally unchanging data such as ship name, ship type, and destination, while dynamic data is locational and time-specific data. We filter these files to ensure that they lie with the specified time and latitude-longitude bounds for the AOR. This limits the overall dataset to just over 17 million observations of dynamic data.

Due to the massive size of the full dataset, we perform the initial data cleaning on a set of 26 tankers. We select them under the condition that they had traveled to the AOR at some point during January 2014. The purpose of cleaning the data is to remove common AIS transponder issues that occur in the data. Another possible problem with the data, which occurs within both the small dataset and full dataset, is two ships having the same Maritime Mobile Service Identity (MMSI) number. Due to the difficulty this causes, we set aside the observations for vessels with the same MMSI number from the dataset along with any points identified as a transponder error. Once we confirm that the

dataset is clean, we show how to construct a dataset of stop points for all vessels. The results of this operation yield a dataset of a more workable size.

After determining the stop points, the next step is to construct an inter-point distance matrix for the clustering algorithm to use. The Vincenty formula, which calculates the distance between two points on a spheroid, is accurate for computing distances with spatial data. To reduce computation time, we use the treeClust function to produce our distance matrix. Because the matrix returned by treeClust does not include actual distances as values, we must compare the post-clustering results from the small dataset using the treeClust method to those using the Vincenty method (Buttrey and Whitaker 2016). We do this by using Cramér's V (Crewson 2012) test to measure agreement between the two results. This test measures agreement by comparing which cluster group every observation falls under using the Vincenty method against which group it falls into using the treeClust method. If the Cramér's V value is considered high enough, then the treeClust clustering results can be considered to have a similar degree of accuracy as the Vincenty clustering results.

We choose the Ordering Points to Identify the Cluster Structure (OPTICS) clustering algorithm over multiple available clustering algorithms because it is able to cluster spatial data with shapes and densities, and because it can identify outliers, or points that do not belong in a cluster. The initial attempt to cluster over the entire AOR returned a low Cramér's V value, signifying a low level of agreement. To remedy this, we split the AOR into seven Universal Transverse Mercator (UTM) zones. Because only 607 stop points are in the small sample dataset, two of the zones did not have enough observations to form a tree, but all of the values returned for the rest of the zones suggested strong agreement between the two methods. The assumption moving forward is that the same agreement will hold when scaling up to the full size dataset.

When moving to the full dataset, data storage and computation time become a major issue. The dynamic data for the full month of January 2014 is over 17 million observations, so to clean the dataset we partition the data and then reassemble it upon completion. Once we determine the stop points for the full dataset, we are able to reduce the dataset size to 179,060 stop points.

The treeClust method did not cluster as well as expected upon visual inspection of the clustering results. While plotting the cluster centers over their respective stop points, it becomes clear that some clusters spanned four or five times the distance of others in order to include enough points to make up a cluster group. For this reason, we then cluster by zone, using UTM northing and easting coordinates. The visual inspection of these clustering results demonstrate that clustering with OPTICS based on UTM coordinates by zone yield reduced computational time while producing reasonable clusters.

Finally, we illustrate how one might train a simple tree model from the clustering results to classify a new observation into the appropriate cluster. We begin this step by partitioning the data using an 80% to 20% split on the training and test set, respectively. Using a classification tree, and pruning it to increase accuracy while simultaneously reducing complexity, we predict the values for our test set. In this case, we use the cluster group found in the previous step as the response variable. The objective for this is to calculate the misclassification rate for each zone as a performance metric. The classification tool could serve as the framework for a threat detection system by comparing the clustering results to the predicted results.

References

Buttrey S, Whitaker L (2016) treeClust: An R package for tree-based clustering dissimilarities. *The R Journal*. 7(2):227-236.

Crewson P (2012) *Applied statistics handbook* (AcaStat Software, Winter Garden, FL).

International Maritime Organization (2017) AIS transponders. Retrieved 19 July, http://www.imo.org/en/OurWork/safety/navigation/pages/ais.aspx.

THIS PAGE INTENTIONALLY LEFT BLANK

ACKNOWLEDGMENTS

I would like to thank Dr. Whitaker for the guidance she provided during the entirety of the thesis process. Without her help, I know that I would have been lost along the way. Also, I offer my gratitude to Dr. Koyak, who agreed to come onboard very late in the process, for his work as my second reader.

I also would like to acknowledge Ms. Megan Guidi for her emotional support and consistent reminders of project deadlines.

Finally, I would like to thank all of the friends and family members who have supported me throughout my time at the Naval Postgraduate School, especially the boys at 11 Portola.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

Safe navigation of the world's waterways is an objective that humans have attempted to achieve since the beginning of time. Given the current state of satellite technology and dedicated maritime support services within various maritime organizations, this goal has largely been achieved in modern times. The Automatic Identification System (AIS) is a systematic approach to monitoring vessels. In 2000, the International Maritime Organization (IMO) mandated the use of AIS as a collision-avoidance system (IMO 2017). Each ship with AIS transmits frequent messages, some giving "dynamic" information such as location, heading, and speed, while others give "static" information such as call sign and vessel type. Because raw AIS messages are now collected and stored in large databases, they can also be used for purposes other than collision avoidance. In particular, they are used for anomaly detection and projecting ship trajectories (Pallotta et al. 2013). A computational task that is often a prerequisite for this type of work is to identify clusters of common stop points. Our focus is to identify clusters of stop points in a way that can be easily scaled to very large datasets of AIS, and be used in regions that have complex shipping behaviors.

A. AUTOMATIC IDENTIFICATION SYSTEM OVERVIEW

Ships are distinguished as Class A regulated vessels and Class B non-regulated vessels. The requirements for a vessel to be outfitted with a Class A AIS device are set by the United States Coast Guard within United States waters, and are as follows:

1. Any self-propelled vessel exceeding 1600 gross tons;
2. A self-propelled vessel of 65 feet or more in length, engaged in commercial service;
3. A towing vessel of 26 feet or more in length and more than 600 horsepower, engaged in commercial service;
4. A self-propelled vessel that is certificated to carry more than 150 passengers;

5. A self-propelled vessel engaged in dredging operations in or near a commercial channel or shipping fairway in a manner likely to restrict or affect navigation of other vessels;

6. A self-propelled vessel that is engaged in the movement of certain dangerous cargo, including flammable or combustible liquids (Department of Homeland Security [DHS] 2017a).

The requirement for vessels that do not meet the conditions for use of a Class A device to operate using an AIS Class B device applies to many fewer vessels, the main condition being that they are not subject to pilotage by other than the vessel master or crew. This applies to fishing industry vessels, vessels carrying less than 150 passengers, and vessels engaged in dredging operations (DHS 2017a). Due to the fact that Class B vessels are engaged in lighter commercial or leisure activities, their devices transmit information at less frequent intervals than their Class A counterparts (DHS 2017b). Although it is important to understand the differences between the two device classes, the focus of this thesis will be to analyze those vessels with Class A devices.

In order for AIS to perform its function successfully, an intricate system of satellites, vessel-based transmitters, and ground operators work together in unison. Each vessel is outfitted with one Very High Frequency (VHF) transmitter and multiple VHF receivers. The system transmits various factors about the vessel autonomously and continuously, while simultaneously checking its transmission schedule to avoid interference from other vessels. In addition, it schedules future transmission slots (DHS 2017c). This process, called Self-Organized Time Division Multiple Access (SOTDMA), allows for a practically unlimited system capacity.

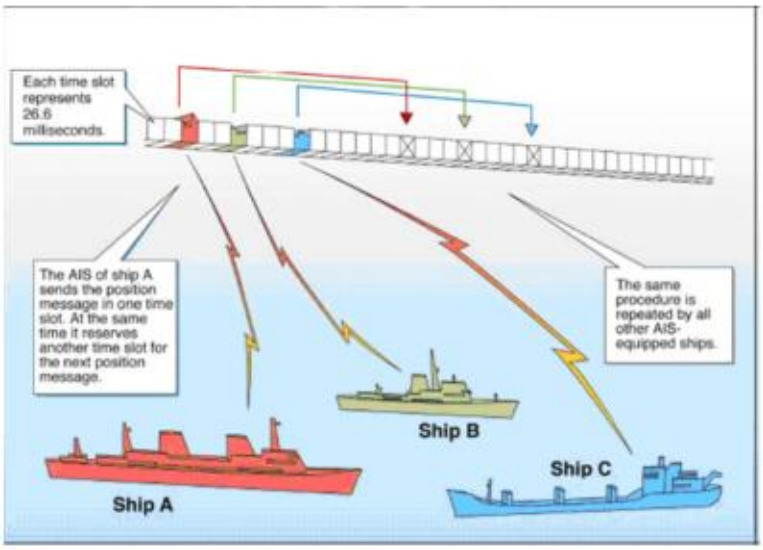

Figure 1. SOTDMA Use by AIS. Source: DHS (2017c).

A vessel's AIS transmissions are received in two ways, as ship-to-ship communications and as repeater transmissions. If a vessel is within 10 nautical miles (NM) of a vessel transmitting AIS, the AIS transmission is sent directly to the nearby vessel using VHF (DHS 2017c). The main limitation to VHF transmissions is that their propagation is limited by the height of the antennae. In order to remedy this issue, certain "repeater stations," such as buoys, have been set up to allow Vessel Traffic Services (VTS) to extend their range in order to access information on vessels entering port areas (DHS 2017c). AIS capabilities have been installed on satellites beginning in 2008, but only recently has usage of this capability become substantial (Strauch 2009). Initially, these satellite-based AIS (S-AIS) systems were useful only in open-ocean regions to provide vessels with an extended range, as VHF signals can propagate vertically much farther than they can horizontally (Ginesi 2009). Recently, certain companies have been launching special satellites into orbit with the intention to create a worldwide S-AIS system (de Selding 2015).

B. AREA OF RESPONSIBILITY

Although AIS is used by vessel and VTS centers across the world, the focus of our study, like the recent work of Bay (2017) who focuses on ship traffic in the Port

Fourchon, Louisiana region, is on ship traffic in the Gulf of Mexico and Caribbean Sea during the month of January 2014. Figure 2 shows a map of the AOR of our study.

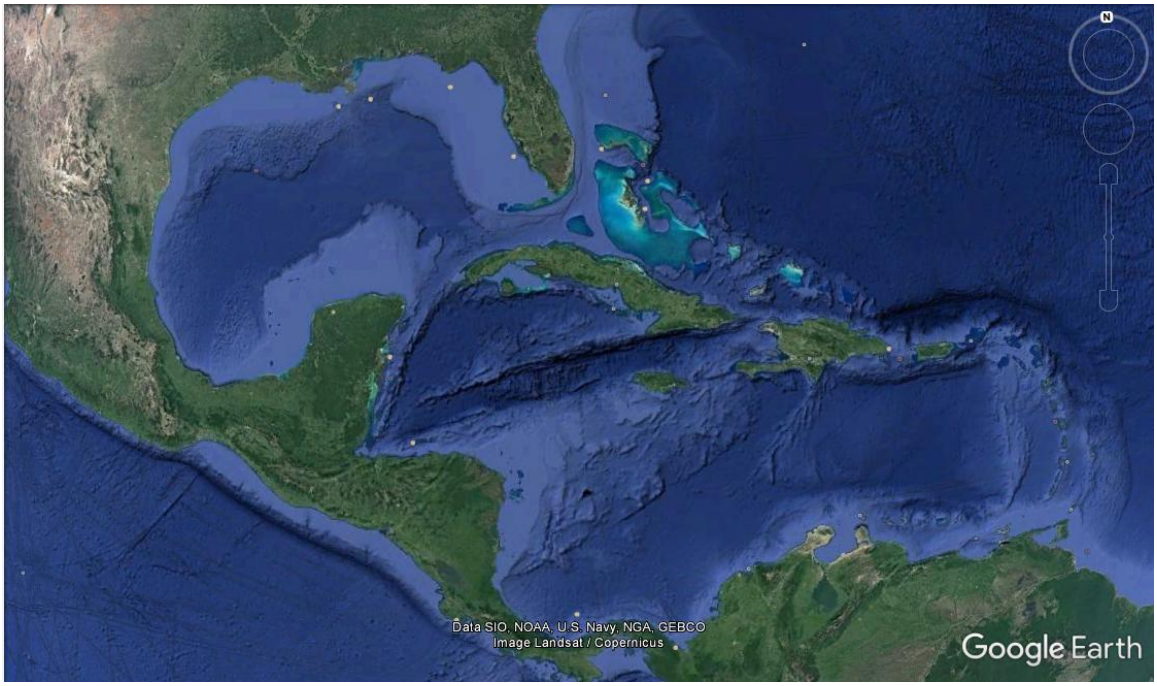

Figure 2. Geographic Representation of the AOR. Source: Google Earth (2015).

This region, consisting of 1.6 million square miles, includes many tropical islands and is an area of substantial maritime economic activity. Tourism, fishing, oil production, and shipping constitute the main forms of economic activities that take place in these waters, making up a total of $234 billion dollars per year (Hargreaves 2010). Ships that operate in this area exhibit navigation patterns that are categorized by their economic activity. Figures 3 through 6, which are screenshots of a "live" map from the website marinetraffic.com, illustrate the behavioral patterns exhibited by vessels of different vessel types (MarineTraffic 2017).

Figure 3 shows that cruise ships and vessels that rely on tourism as their source of income tend to travel along routes that stay close to the shoreline. For Figures 3−6, a circle represents a stopped vessel, while an arrow represents a vessel underway.

Figure 3. Passenger and Pleasure Craft Behavior. Source: MarineTraffic (2017).

In contrast, Figure 4 shows that fishing vessels generally display one of two behaviors, which are largely dependent on vessel size. Larger vessels tend to journey out and spend multiple days farther from shore as they make their catch. Smaller craft tend to travel less and make their catch, then return to their homeport in the same day. These small "onshore" boats numerically dominate their larger offshore counterparts (NRC 1991). A shortcoming of Figure 4 is that it only accounts for vessels tracked using AIS, and a majority of fishing vessels do not use AIS. A fishing vessel that does not meet the previously described requirements would not appear on the tracker, which mostly includes the small, single-day trip vessels.

Figure 4. Fishing Vessel Behavior. Source: MarineTraffic (2017).

Figure 5 shows how cargo vessels move in the AOR, exhibiting a high traffic density. It also shows somewhat defined lanes or routes used by these large cargo vessels. Although there is a constant, substantial flow of shipping through the AOR, commercial shipping and fishing combined constitute only about one percent of its total economic activity (Hargreaves 2010).

Figure 5. Cargo Vessel Behavior. Source: MarineTraffic (2017).

The behavior displayed by oil tankers is very similar to that of cargo vessels, but with a significantly higher concentration of vessels off the coast of Texas and Louisiana. This may be due to the 6,364 offshore platforms located within the AOR (Bay 2017). The oil and gas industry makes up 53% of the economy for the AOR, and will continue to increase its role as North America pushes toward its independence from foreign oil (Hargreaves, 2010). These behaviors are important to consider throughout the analysis, as they are somewhat specific to the AOR.

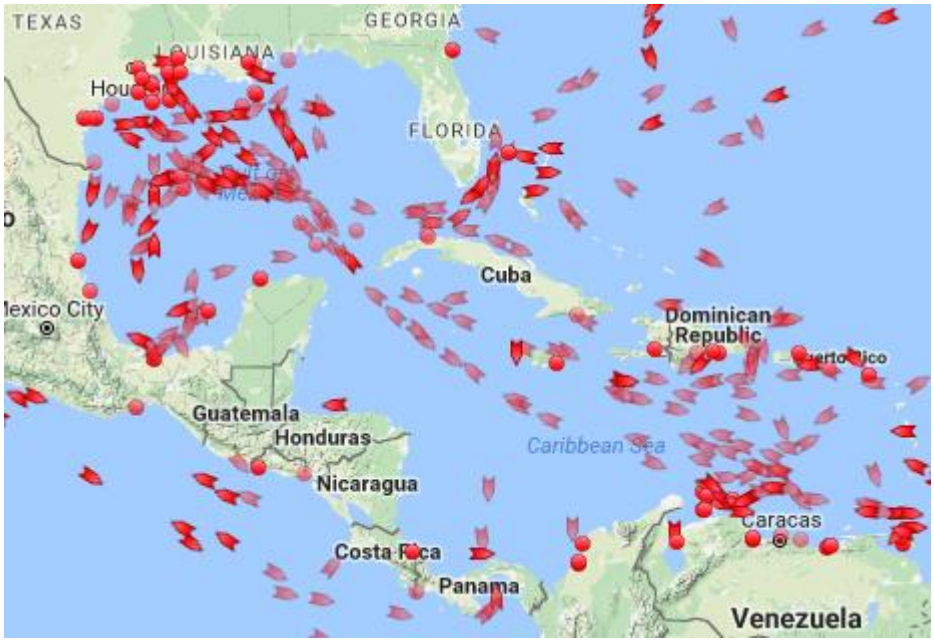

Figure 6. Oil and Gas Tanker Behavior. Source: MarineTraffic (2017).

C. RESEARCH OBJECTIVES

We focus on stop points rather than on moving ships or ship tracks. While most work with AIS data focuses on detecting anomalous behavior of moving ships (e.g., Mao et al. 2016), the behavior of stopped ships is important. Activities such as staying overlong at a particular stopping region such as a fishing area or port might indicate suspicious behavior. Another indication of suspicious behavior might be a sequence of stopping regions that is unusual for that vessel or for vessels of the same type. Our work sets the stage for investigating and identifying these types of anomalous behaviors.

Our study has two main objectives. First, we aim to find an efficient method to transform the month-long collection of AIS data with over 17 million observations into a usable, well-organized data set that might be used for identifying stop points or for other AIS related work. This is important, as the Center for Maritime Research and Experimentation (CMRE) received 600 million AIS messages per month from various sources as of the year 2013 (Pallotta et al. 2013) and the rate has continued to increase as more vessels use AIS. We note that Pallotta et al. (2013) proposes a scheme to convert AIS data from a completely raw format into a maritime movement database. The authors

develop an automated method to detect behavioral anomalies, and to predict the motion of vessels. Their algorithm, called Traffic Route Extraction and Anomaly Detection (TREAD), analyzes vessels as a collective entity, and uses their behavior as a standard for its low-likelihood behavior detection algorithm. Our study differs from that of Pallotta, et. al (2013) in that our data cleaning efforts, while designed to study stop points, are steps that also need to be taken when using AIS data to study moving ships.

Our second objective is to use this investigation as a starting point for clustering vessels into stopping regions using geospatial AIS stop-point coordinates along with additional ship information such as vessel type and size, its voyage and stopping history, and any other available information. There are several issues with clustering AIS stop points. The first issue has to do with choice of clustering algorithm. Stopping regions have different shapes and different densities of AIS messages. Some are long and narrow while others are diffuse and cover large areas. The second issue has to do with measuring inter-point distances between observations that contain both geospatial coordinates and other possibly mixed-type categorical and numeric variables. We study the feasibility of using the treeClust algorithm of Buttrey and Whitaker (2016) to compute inter-point distances between stop points combined with the density based clustering algorithm Ordering Points to Identify the Clustering Structure (OPTICS) of Ankerst et al. (1999). The OPTICS algorithm will identify clusters of different shapes and densities and can identify outliers.

Finally, because the inter-point distances of Buttrey and Whitaker (2016) are learned for a particular dataset, and because computation of inter-point distances and clustering is computationally intensive, we show how the results of clustering AIS might be used to train a classification algorithm that can be used quickly in real-time to identify cluster identities of new AIS stop points as they arise.

D. THESIS ORGANIZATION

The remainder of our thesis is organized in the following way. In Chapter 2, we describe the data contained in AIS messages. In Chapter 3, we explain the methodology that we use to clean the data, convert the dynamic data into stop points, cluster the stop

points, and create our classification tool. In Chapter 4, we present the results from our analysis, and Chapter 5 contains the conclusion and recommended topics for future research.

II. BACKGROUND

The CMRE is receiving AIS data at an ever-increasing rate (Pallotta et al. 2013). The raw messages that are both sent and received follow specific guidelines established by the National Maritime Electronics Association (NMEA), and are in a format that must be decoded in order to be usable. We use software from the open source library "libais," written in C++, to improve the speed of decoding the raw AIS messages (Schwehr 2017). The result is a database that includes worldwide AIS data in a readable JavaScript Object Notation (JSON) format. The decoded records are then converted to a Comma Separated Value (CSV) file that contains approximately 12 million messages a day, covering the four-month period of January through April of 2014. Limiting the decoding to include only Class A vessels, and removing observations that cannot be properly decoded drops the total file size to about 8 million messages a day. Most of these observations come in the form of two distinct formats known as "static" and "dynamic" messages. We describe each of these formats below.

A. STATIC MESSAGES

Class A static "category 5" messages contain information that describe a single ship on a single voyage. A non-automated static report must be broadcasted every 6 minutes by the vessel (DHS 2017d). The information included in a static message is described in Table 1.

Table 1. Information Fields for Static Data

Field Name	Type	Description
Maritime Mobile Service Identity (MMSI)	9-digit Integer	Valid MMSI have first three digits between 201 and 775, although there are a few exceptions.
International Maritime Organization (IMO) number	7-digit Integer	Unique hull number. 0 not available; 0001000000 – 0009999999 valid IMO
Radio Call Sign	Character	Free form text.
Ship Name	Character	Free form text, maximum 20 characters.
Reference Point or Ship Dimension	Numeric (m)	Four fields giving the distance to the reference point from port (C), starboard (D), bow (A) and stern (B). If C = D = 0, the A and B give length and width. Maximum value for A, B is 511m. Maximum value for C and D is 63m .
Destination	Character	Free form text, maximum 20 characters.
Estimated Time of Arrival (ETA)	Character	UTM in YYDDHHMM format.

The first three digits of a ship's Maritime Mobile Service Identity (MMSI) number are known as the Maritime Identification Digits (MID). This three-digit code is used to determine the ship's nation of origin. MMSI, IMO number, radio call sign, and ship's name all serve to identify which ship is transmitting the AIS message. The IMO number is the most reliable identifier of a vessel, as it is unlikely to have two vessels with identical IMO numbers, which may not be the case with the other ship identification fields. A shortcoming is that not all vessels are required to have an IMO number (IMO 2017). Another important aspect to mention is that both the destination and Estimated Time of Arrival (ETA) fields are entered by the ship's crew, which raises the possibility of errors.

B. DYNAMIC MESSAGES

Class A dynamic messages are message types 1, 2, and 3, and are transmitted much more frequently than are their static counterparts. They must be transmitted every 2-10 seconds while underway, and every 3 minutes while at anchor (DHS 2017d). Consecutive messages may be plotted in order to track ship movement over time. Table 2 shows the different data fields that are included with these types of messages.

Table 2. Information Fields for Dynamic Data

Field Name	Type	Description
MMSI	9-digit Integer	The key used to pair dynamic and static records, see Table 1.
Navigational Status	Integer	Valid codes are between 0 and 15.
Rate of Turn	Integer (degrees per min)	Valid values are between -127 and 127 with 0, negative, and positive values indicating no, left and right turns respectively; and +127, -127 being the maximum reported turn rate; -128 indicates no turn information. But other values are observed.
Speed Over Ground (SOG)	Numeric (0.1 knots)	1022 is 102.2 knots or higher, 1023 is a missing value
Course Over Ground (COG)	Integer (0.1 degrees from North)	Valid values should be between 0 and 35999. 3600 indicates missing value. (larger values are observed).
Latitude and Longitude	Numeric (degrees)	Position. Latitude of 181 and Longitude of 91 indicate missing values.
True Heading	Integer (degrees from North)	Valid values should be between 0 and 359, but larger values are observed, and 511 indicates a missing value.

Field Name	Type	Description
Time Stamp	2-Digit Integer (seconds)	The time in seconds 0 – 59 with missing values indicated by 61, 62, 63.
Time	Integer (seconds, UTM)	Time since some reference point. This is an additional field, not part of the message payload.

Although it is somewhat unintuitive, the time field is not used to show the time of transmission, but rather to determine the possibility of radio interference. A field for time of transmission, the time stamp field, is available but must be included as an addition to the message. The time stamp is reported in universal transverse Mercator (UTM) units, which must be converted for use in analyses that require local time. The dynamic data are also subject to various errors. The first two fields, MMSI and navigational status, may be entered incorrectly by the ship's crew. For example, in our data we see MMSI 123456789 assigned to what appear to be several different vessels. Errors are due to instrumentation and transmission. There are cases where ships exhibit high speeds that are physically impossible for maritime vessels, which may be due to errors in location (GPS) or in the time stamps. These cases must be identified and removed from the dataset before any analysis is attempted.

III. METHODOLOGY

A. DATA CLEANING

In our research, we consider AIS data collected for the period of 1−31 January 2014. This dataset has over 17 million observations when limiting the AOR by a latitude band from 8° to 31° and a longitude band from -98° to -58°. We illustrate our data cleaning steps, and how we identify stop points on a much smaller subset of the data. We use a selection of 26 tankers that traveled into the AOR during January 2014. This helps to save computation time, and does not affect our function's efficiency when scaling up to the full dataset

1. Transponder Issues

One of the major issues with AIS messages is that there is seldom a case where no errors occur during a trip. The three most common instances of errors are duplicates, teleportation, and infeasible speeds. Duplicates are by far the most common, in that different messages contain identical values in all fields. Teleportation occurs when a vessel appears at a latitude and longitude completely off course for a single message, and then regains its track. We determine if a point has teleported by comparing the latitude and longitude to the time. If two points with the same MMSI during the same time have a differing latitude and longitude pairing, then the point off of the vessel's track is marked. As shown in Figure 7, where different ship tracks are indicated by different colors, there are single teleported points that appear with no "connecting" points in sight. These anomalous points are circled within Figure 7. Finally, we identify infeasible speeds after calculating the latitude and longitude change compared to the change in time. This calculation flags vessels travelling faster than 60 knots.

All computation, including these data cleaning steps, are performed using the statistical computing software R (R Core Team 2017) as implemented in RStudio (RStudio Team 2016). Our R data cleaning function serves two distinct purposes, the first of which is to create a vector that marks whether or not an observation has one of the three common error types. Once it identifies which observations are errors, it calculates

the distance in nautical miles, speed in knots, and time change in seconds between consecutive error-free messages. By identifying messages containing errors, we are able to remove them from the dataset. However, we postpone this step due to another less common, but significant issue within the data.

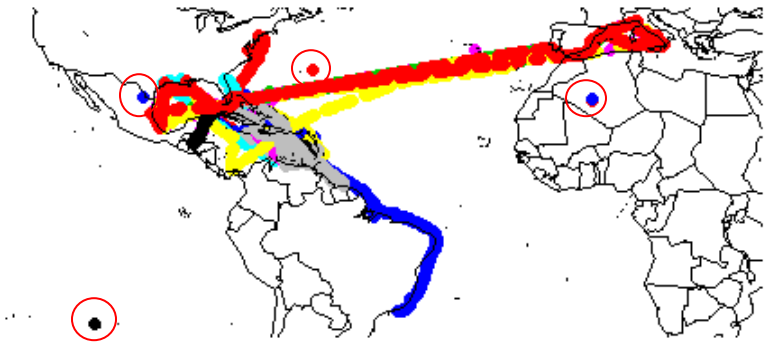

Figure 7. Plot of Uncleaned Message Points by MMSI Number Showing Teleportation

2. Ships with the Same MMSI Number

The issue of two ships containing the same MMSI number is not very prevalent, but can occur from time to time, and indeed one of the MMSI numbers among the 26 in the small subset of tankers belongs to two ships. We know this by inspecting both static and dynamic records, which reveal two IMO numbers and ship names. MMSI numbers are granted by local authorities, and an accidental duplication is possible. In order to remedy this issue, we develop an algorithm that identifies if it is likely that two vessels have the same MMSI number based only on dynamic records. We determine whether two vessels have the same MMSI number based on their "Overall Error Level" (OEL). We calculate OEL by dividing the number of messages, including any of the three error types, by the total number of messages transmitted with a given MMSI number. This returns a value of about 5% for 24 vessels in the small subset of tankers, but for the two vessels with the same MMSI number, the OEL is close to 60%. Using an OEL threshold of 30%, our function returns a list of MMSI numbers that most likely belong to more than one vessel when given a dataset.

While in some cases it may be possible to resolve this issue by creating vessel identifiers from a combination of IMO numbers and MMSI numbers, the simplest solution is to remove these vessels from the dataset. After completing this step, it is now possible to set aside all of the messages containing errors. Figure 8 shows a map of the points with only 25 MMSI numbers and no messages containing errors. Notice there are no longer single, stray message points, as compared to Figure 7.

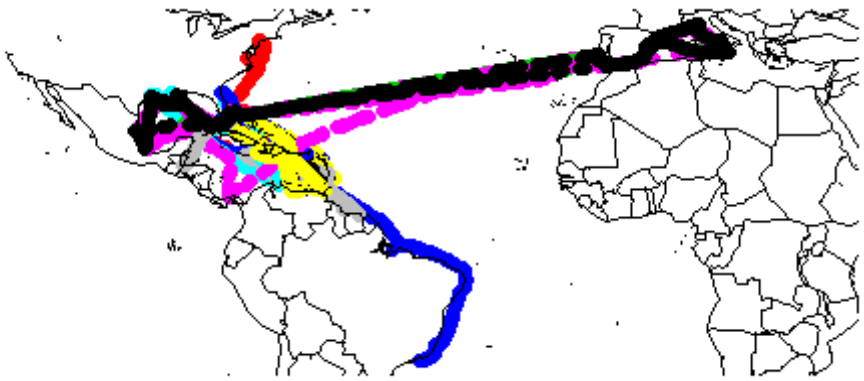

Figure 8. Plot of Cleaned Message Points by MMSI Number

While an OEL threshold of 30% may not be the best choice globally for both identifying ships with the same MMSI number and for reducing the false positive rate for those that do not, it is unlikely that ships travelling in similar patterns during the same time period have been appointed the same MMSI number.

3. Determination of Stopping Points

The next step in handling the data is to determine which of the messages were sent at times when the ship was stopped. This is a difficult task for many reasons. First, while a vessel is anchored it is possible for it to move around with the wind and current for it to register a change in its coordinates. Second, vessels will often slow or stop in high traffic areas in order to maintain safe practices in collision avoidance. Finally, a vessel, such as a trawler, may travel in a pattern that causes it to transmit its position at the same time in its motion path, causing it to appear as if it had stopped. Due to the

many examples where an error in detecting or misclassifying a stop point can occur, we make the following assumptions:

1. Any vessel travelling at a speed less than 0.3 knots should be considered stopped, as no vessel would be travelling at this speed to perform a maritime activity.

2. If a ship moves for less than 30 seconds between points that have been determined as stop points, then the travel messages will be disregarded.

3. Any stops that are less than 5 minutes in length are not long enough to perform an operation, such as refueling, fishing, etc., and should not be included in the final stop point dataset.

4. It is understood that constant wind and water movement will cause a slight changes in a ship's positioning, so we represent the stop point by the average latitude and longitude of the points at a single stop.

Using these assumptions, we develop an algorithm that requires the MMSI, time stamp, and location of a cleaned dataset; and returns a dataset consisting of the MMSI number, average latitude/longitude, and total time stopped for each step. For the tanker subset, we identify 607 stop points from the original 101,000 records. Figure 9 shows the location of these stop points, colored by their vessel's MMSI number.

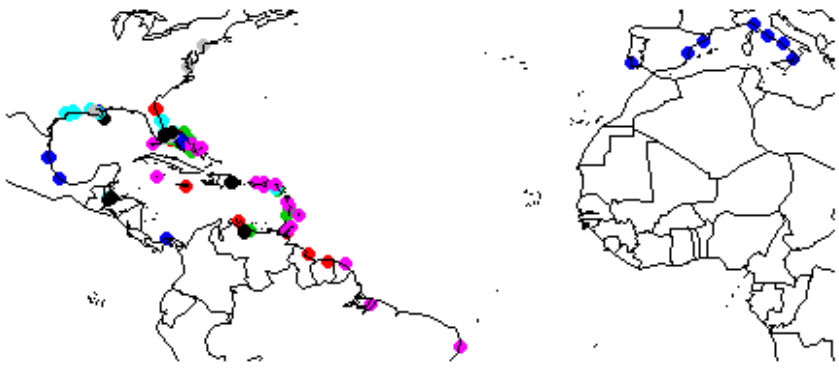

Figure 9. Plot of Stop Points by MMSI Number

As seen in Figure 9, there is a very high concentration of vessels stopping in what seems to be Port Canaveral, FL, as well as Port Fourchon, LA. It is also clear that further clustering of these stop points is required to identify regions where vessels tend to stop, such as ports, fishing areas, or offshore oil platforms.

B. CLUSTERING METHODS

The clustering algorithm we use on the dataset of stop points is OPTICS (Ankerst et al. 1999), as it is implemented by the function `optics` from the R package `dbscan` (Hahsler and Piekenbrock 2017). This algorithm is an extension of the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm (Ester et al. 1996). The DBSCAN algorithm uses a set radius and a set minimum number of points to identify core points, or points that fall in a cluster. A point is a core point so long as it meets the threshold for the minimum number of points within the set radius. All points that are not core points are classified as outliers, and placed into a cluster group 0. DBSCAN is a density-based algorithm. Using a density-based algorithm for identifying stop-point clusters has two advantages over using other more well-known partitioning algorithms such as K-Means (see MacQueen [1967]). The first is that density-based algorithms identify high-density regions of any shape, such as long narrow stop-point regions along a river bank, or small circular stop-point regions around an oil platform. Secondly, density-based algorithms do not force cluster membership on every point in the dataset. Observations that do not have the required minimum number of points within the specified radius are allowed to be outliers or noise. This facilitates identifying potential anomalies and makes the method more robust to transmission errors not accounted for in the cleaning steps. Finally, the DBSCAN algorithm can be scaled to cluster very large datasets because it is fast, parallelizable, and amenable to distributed type computation (Ester et al. 1996).

While DBSCAN is a robust tool for clustering, a major weakness is that it struggles to identify clusters of varying density (Ankerst et al. 1999). This is particularly important because the AOR contains both low-density stopping regions, such as fishing areas, and high-density stopping regions, such as the Port of Miami. The OPTICS

algorithm is able to compensate for this by first ordering points individually so that points near each other in this ordering are in the same cluster. The ordering can be viewed in a special type of dendrogram known as a "reachability plot." For a more thorough definition of reachability, see Ankerst et al. (1999). As an example, Figure 10 displays the reachability plot where distances between points are computed using the Vincenty method (Karney 2013), as discussed in the next section.

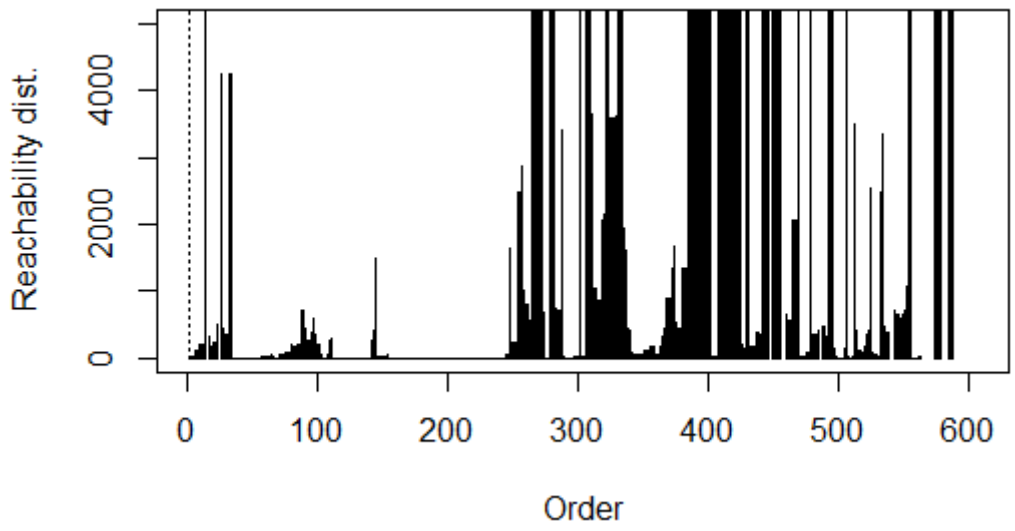

Figure 10. An OPTICS Reachability Plot Example

The valleys in between spikes in the reachability plot are points that belong to the same cluster group. For example, we see a cluster of points at about point 350 on the horizontal axis of Figure 10. Furthermore, by looking for valleys between lower spikes, one can identify a hierarchy of clusters. In Figure 10, there appears to be a large cluster of points 25 through 275 that contain subclusters of points 25−80, 100−150, and 150−275. The main benefit of using this type of plot is that one can visually determine how the clusters are to be grouped, based on setting a threshold reachability distance. Once we make this determination, it is possible to extract the cluster groups and identify outlier points that do not belong to any cluster.

1. Vincenty Inter-point Distance Method

We cannot utilize the full stop-point dataset directly as input for `optics` because when given a dataset `optics` computes inter-point distances as Euclidean distances. This technique does not accurately compute distances for geospatial data with latitude and longitude coordinates. The `optics` function can use other distances, but the user must compute those distances and provide `optics` with an inter-point distance matrix as input in place of the dataset. For small datasets, such as the dataset with 607 stop points, we construct an inter-point object of class "dist" using Vincenty distances. While using this approach provides a simple technique for using the `optics` function, the issue comes when scaling up the size of the data. For a dataset of size *n*, the calculation of the inter-point distance matrix is on the order of $O(n^2)$, which for large *n* requires a large amount of computational time and memory. For this reason, and to have a method that includes other clustering variables, we compare clustering results found using the treeClust distance method, discussed in the next section, to the clustering results found using the Vincenty method. The metric of performance used to compare the two methods is the value for Cramér's V (Crewson 2012), which measures agreement between two clusterings. Because we already know that the Vincenty method results are accurate, then the Cramér's V value will tell us if the comparative method is also providing accurate results.

2. treeClust Method

The treeClust method of Buttrey and Whitaker (2016) implemented in the R package `treeClust` (Buttrey 2016) uses classification and regression trees to "learn" inter-point distances. We use d3, the third of four options for computing treeClust distances. Rather than use the inter-point distance matrix, which for large datasets would overwhelm R's memory, we map the data to Euclidean space in a way that tries to preserve the treeClust inter-point distances. See Buttrey and Whitaker (2016) for details. Inter-point distances between observations in the resulting "newdata" dataset can then be computed as Euclidean distances. Thus the "newdata" dataset can be used as the input to the `optics` function.

The advantages for using this method go much farther than reduced computation time. It also allows for the inclusion of additional explanatory variables, including non-numeric variables, into the inter-point distance matrix. This could potentially provide additional insight and improve the accuracy of clustering results.

3. Clustering by UTM Zone

To improve the accuracy of clustering results further, we partition the stop points by UTM zone, and then cluster points in each zone separately. By dividing the dataset into seven 6° longitude bands corresponding to the seven UTM zones that cover the AOR, clustering with the previously described treeClust method, and combining the results; we are able to have a smaller margin of error due to the smaller area for each zone. Figure 11 shows how we divide the AOR into seven zones, UTM zones 14 through 20.

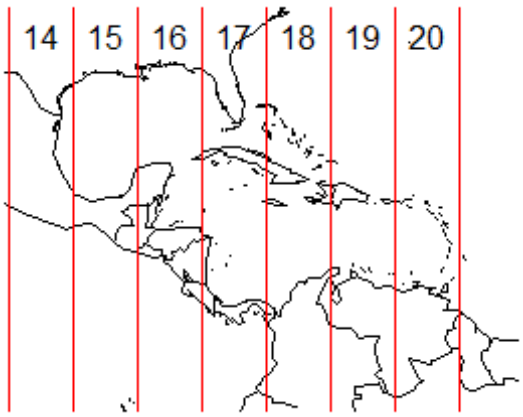

Figure 11. UTM Zones in the AOR

Figure 11 shows that while the zones are evenly divided among longitude bands, the area where a vessel could potentially stop is very different for each zone. In this case, we compare the results of Cramér's V test from the Vincenty and treeClust methods for each zone.

It is possible to project the geospatial coordinates of points in a UTM zone into UTM northing and easting coordinates. With UTM projections, the Euclidean distance

between points in the same zone is approximately equal to the actual distance between the points. Karney (2011) discusses the accuracy of UTM projections, noting that they are most accurate near the equator and meridian of the zones. For our purposes, if we wish to use the `optics` function to cluster a large number of stop-points based only on location, we would first transform the geospatial coordinates to UTM northings and eastings for each zone. We would then use each zone's dataset as input for the `optics` function.

C. SCALING UP THE DATASET

As discussed in the Background section of this paper, the full dataset contains one month's worth of worldwide AIS messages. We note that importing all January 2014 AIS points in the AOR in a single step exceeds default RStudio memory limits. Thus, we clean and identify stop points for the dataset a piece at a time. Although, in this instance, we do computation on each piece in sequence, it would be a simple matter to distribute computations over multiple cores. Once we determine the stop points for the full dataset, memory limits are no longer an issue, as we are able to reduce the total dataset size to 279,860 stop points. As a final test for clustering results, we visually inspect clusters within each zone to ensure that there are no blatant errors.

D. CLASSIFICATION TOOL

After clustering the full dataset, we divide it into training and test sets with an 80/20 split among the total number of observations. We choose this ratio from the Pareto principle, which states that many natural phenomena exhibit a relationship where 80% of the output is a direct result of 20% of the input (Kiremire 2011). In following with splitting and analyzing our cluster results by zone, we construct a classification tool by zone. We train a classification tree on the training set using the `rpart` R package (Therneau et al. 2015), and prune it to reduce complexity. The final metric of performance for the study is the misclassification rate for the cluster group of the test set.

THIS PAGE INTENTIONALLY LEFT BLANK

IV. ANALYSIS

A. SMALL SUBSET CLUSTERING RESULTS

The initial step towards choosing the proper methodology is to determine the Cramér's V value when clustering stop points over the entire AOR. We compare the two values with the OPTICS parameter's minimum points set to five and the maximum reachability distance of 10,000 meters for both methodologies. Since the methods return their reachability plots in different units, we extract the cluster group using the mean reachability value as the threshold for each. Table 3 shows the "rule of thumb" for interpreting the values from a Cramér's V test.

Table 3. Interpretation of Cramér's V Values

Value Range	Agreement
0.00 – 0.10[a]	Weak
0.10 – 0.30	Moderate
0.30 – 0.50	Strong
0.50 – 1.00[b]	Practically the same

Adapted from Crewson (2012).

[a]A value of 0 denotes statistical independence

[b]A value of 1 denotes a perfect relationship

After testing the two methodologies, the `cramer` function from the treeClust package (Buttrey 2016) returns a value of 0.269, which is not quite strong enough to have faith that the assumption will hold when scaling up to the full dataset. The next step is to separate the dataset into UTM zones and perform the same analysis in order to determine if it is more successful. It is possible that by setting the zones as strict boundary lines we are producing more outliers. This occurs when a cluster center for a small group falls on the edge of a UTM zone, and the separation causes the number of points in the cluster group to drop below five. Because it is very unlikely that an event such as this would occur, we do not take any preventative action during sorting. The results of clustering within each of the seven UTM zones are shown in Table 4.

Table 4. Clustering Results by UTM Zone

UTM Zone	Cramér's V Value
14	Too few observations (14)
15	Too few observations (19)
16	0.6211
17	0.6989
18	0.4781
19	0.7260
20	0.5516

From Table 4, when clustering by zone most values meet the threshold of agreement where they are practically measuring the same grouping. For the one zone that does not meet this threshold, Cramér's V still falls under the strong agreement category. In zones 14 and 15 there are too few observations to use the treeClust method, so no value for Cramér's V can be generated. These results indicate that we prefer to separate the data by zone when moving forward to the full dataset.

B. FULL DATASET CLUSTERING RESULTS

During the initial data cleaning steps for the full dataset, we find that 11 ships breach the OEL threshold of 30%, so all of their observations were set aside. Figure 12 shows the OEL distribution for the dataset without the 11 vessels that breach the OEL threshold.

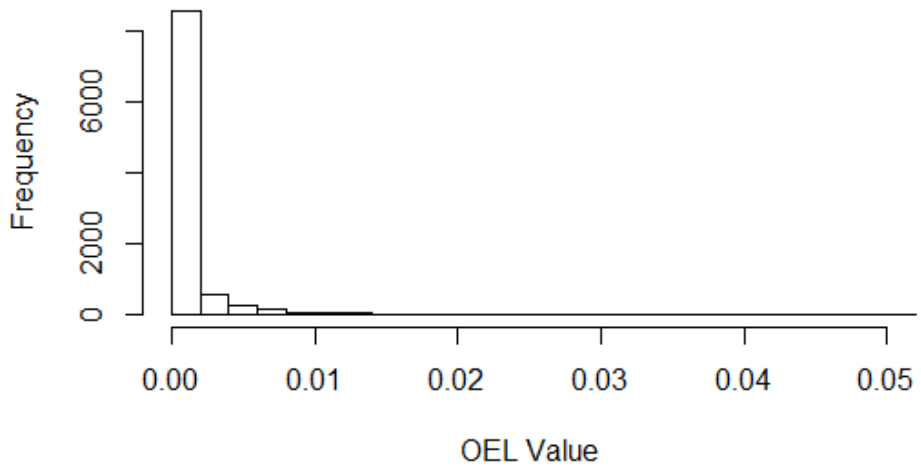

Figure 12. OEL Distribution for Full Dataset

The mean OEL for the full dataset is approximately 0.0016 with a standard deviation of 0.0183. These values show that the 30% threshold can identify two vessels using the same MMSI number. The 11 MMSI numbers that we remove from the dataset have an average OEL of 0.4795, well above the OEL threshold. The next step is to remove the remaining error points. This step removes 24,301 error points constituting 0.142% of the full data set. Figure 13 shows the location of these points within the AOR.

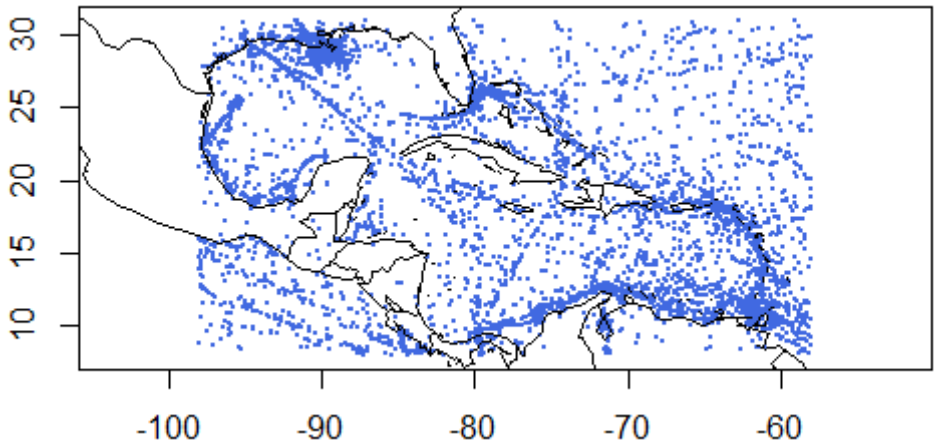

Figure 13. Location of Error Points in the AOR

As shown in Figure 13, these error points are distributed across the AOR, but seem to have higher concentrations in areas that would be ship routes. Having performed the initial data cleaning steps, we generate the stop points. Because the full dataset contains over 17 million observations, we generate the stop points in sections. We take the precaution of splitting the dataset by MMSI number in order to ensure that we do not unintentionally identify unnecessary stops or duplicate stops. This step yields 279,860 stop points over the seven UTM zones. Figure 14 shows the distribution of stop points over the AOR.

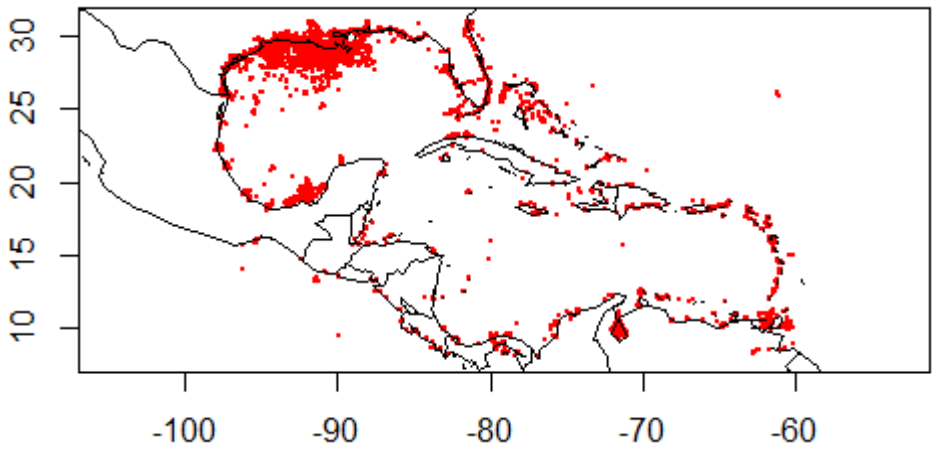

Figure 14. Stop Points in the AOR

As seen in Figure 14, the highest concentration of stop points lie in the area off the coast of Texas and Louisiana. This is most likely due to the fact that 14.0% of all ships in this dataset are tankers. See Bay (2017) for a discussion of shipping and AIS traffic in this area. To extract the number of cluster groups using OPTICS, we start with 0.25 as the reachability threshold. After visually inspecting how the cluster centers compare to the stop points within the zone, we change this value as necessary. Our goal while clustering is to stay consistent, while also forming a reasonable number of cluster groups for each zone. Ultimately, visual inspection shows cluster groups formed based on treeClust inter-point distances using geospatial coordinates by zone, while promising, is

not as accurate as we would like. Figures 15–17 show the location of the cluster centers in relation to their cluster groups using the treeClust method.

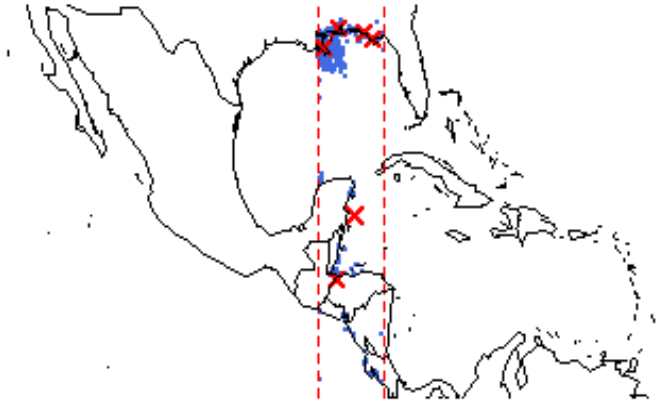

Figure 15. Location of Cluster Centers for Zone 16 Using treeClust Method

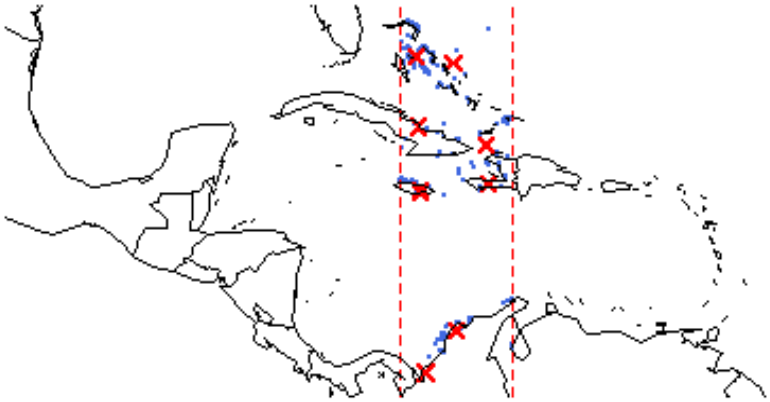

Figure 16. Location of Cluster Centers for Zone 18 Using treeClust Method

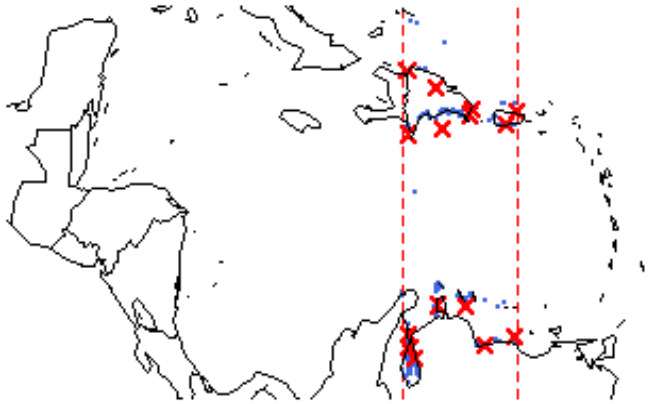

Figure 17. Location of Cluster Centers for Zone 19 Using treeClust Method

Because the treeClust method is scale invariant and robust to monotonic transformations, we are able to use it with the original geospatial latitude and longitude coordinates within each zone. However, because treeClust is designed to take advantage of dependence among variables, it does not perform as well when there are only a few variables (Shaham 2015). The within-zone treeClust clusterings do seem to agree with the Vincenty clusterings, suggesting that this approach might be promising when there is more information available for each of the stop points. The inclusion of additional static data variables, such as those describing ship type, cargo and size, as well as variables capturing voyage and stopping history for each MMSI could improve the accuracy of the clustering results when using treeClust by UTM zone.

The next method we analyze is simplistic compared to the others, but contains some obvious limitations. We cluster by zone again, but this time using only northings and eastings that we convert from their original latitude/longitude pairs. Since `optics` computes its own Euclidean inter-point distance matrix from the data, we input the data directly after conversion. We perform this step under another considerable underlying assumption, that Euclidean distance will be sufficiently accurate over a full UTM zone. Table 5 shows the performance metrics we find after clustering over each zone.

Table 5. Clustering by Zone Performance Metrics

Zone	Reachability Distance	Number of Stop Points in Zone	Number of Cluster Groups	Number of Outliers
14	20,000	12,123	9	20
15	30,000	189,615	24	19
16	25,000	36,378	26	38
17	35,000	20,845	26	27
18	20,000	2,527	33	77
19	20,000	7,624	26	33
20	25,000	10,748	35	16

It is clear that the majority of observations lie within zone 15, but there does not seem to be any correlation between the number of observation in each zone and number of clusters. For the majority of the zones, there are multiple small cluster groups. While having many small clusters increases the complexity of the grouping in each zone, it also greatly improves accuracy by ensuring those points are not absorbed by a larger nearby cluster. Figures 18−31 show a plot of cluster centers overtop of stop points, along with a distribution of number of observations by cluster group for each zone, respectively.

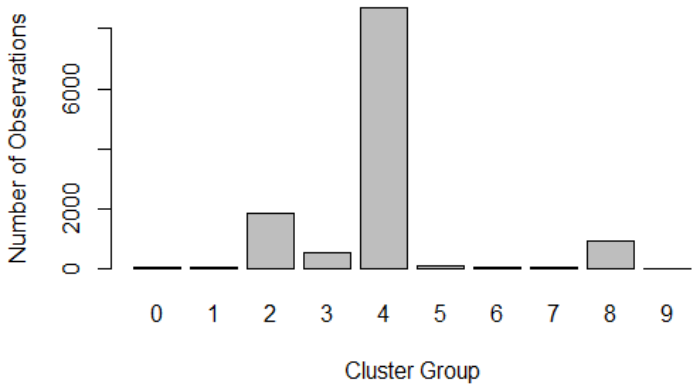

Figure 18. Distribution of Points by Cluster Group in Zone 14

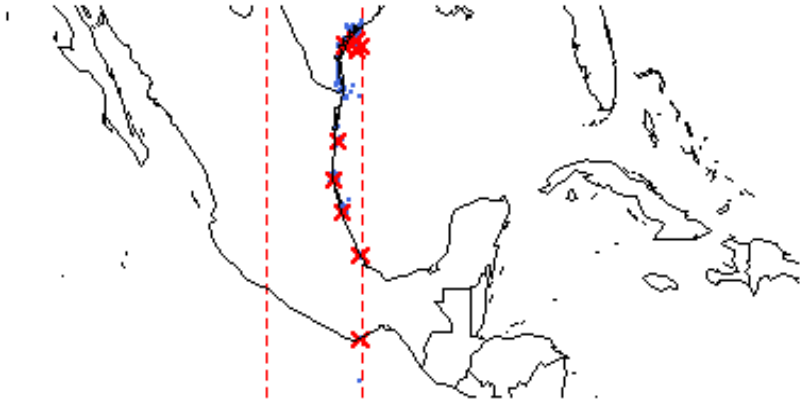

Figure 19. Location of Cluster Centers in Zone 14

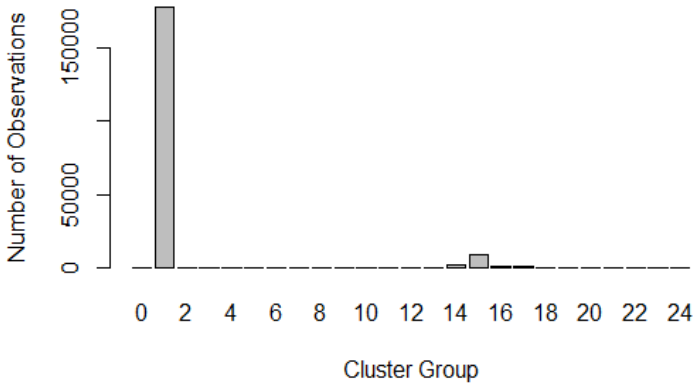

Figure 20. Distribution of Points by Cluster Group in Zone 15

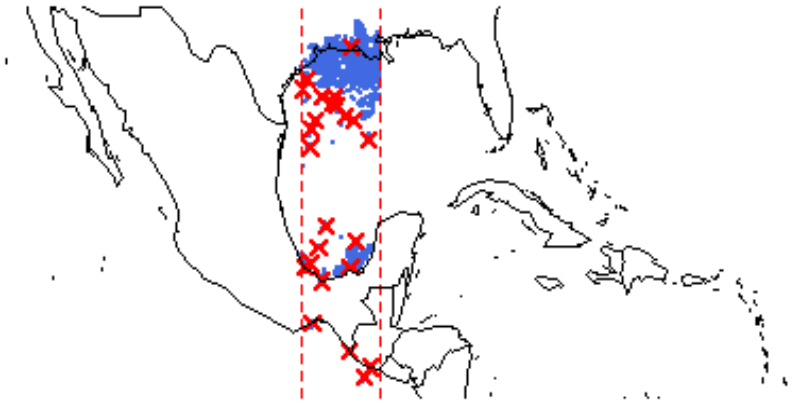

Figure 21. Location of Cluster Centers in Zone 15

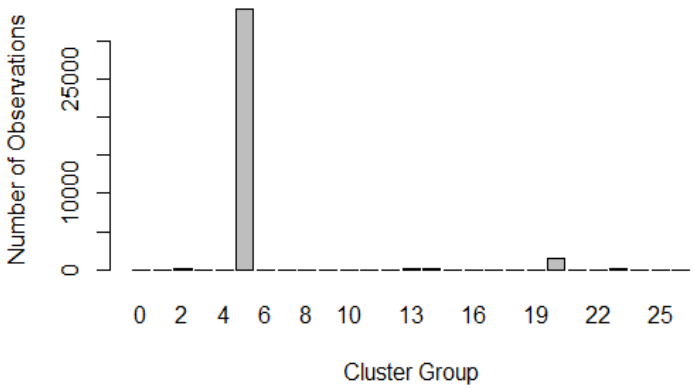

Figure 22. Distribution of Points by Cluster Group in Zone 16

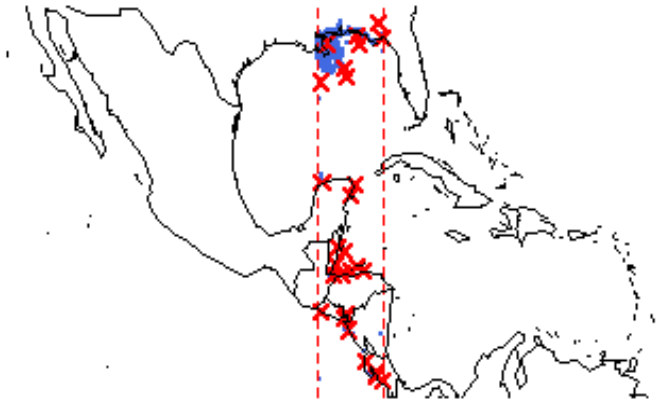

Figure 23. Location of Cluster Centers in Zone 16

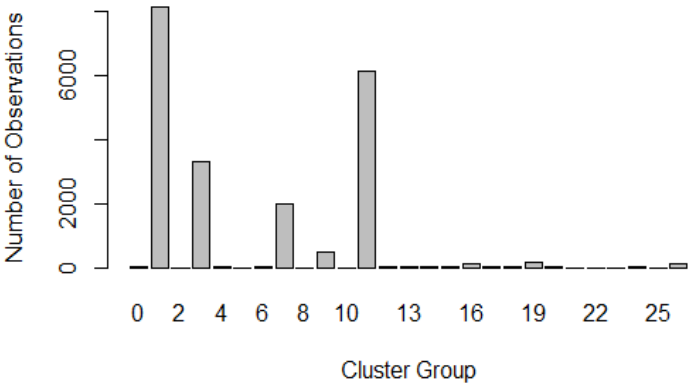

Figure 24. Distribution of Points by Cluster Group in Zone 17

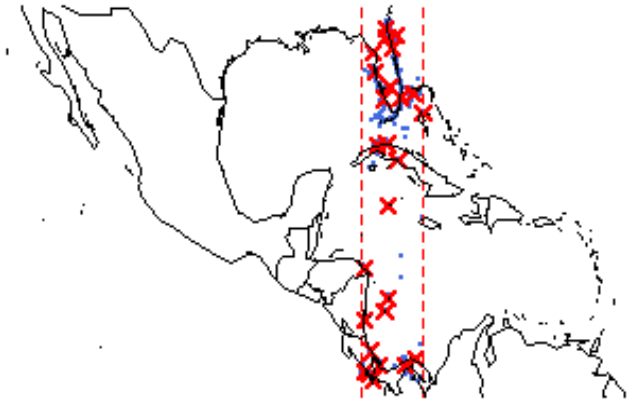

Figure 25. Location of Cluster Centers in Zone 17

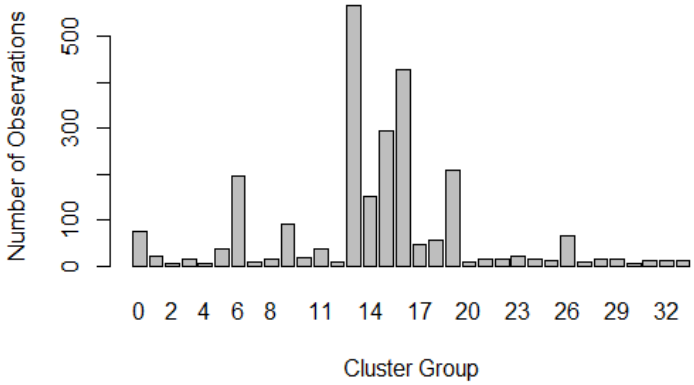

Figure 26. Distribution of Points by Cluster Group in Zone 18

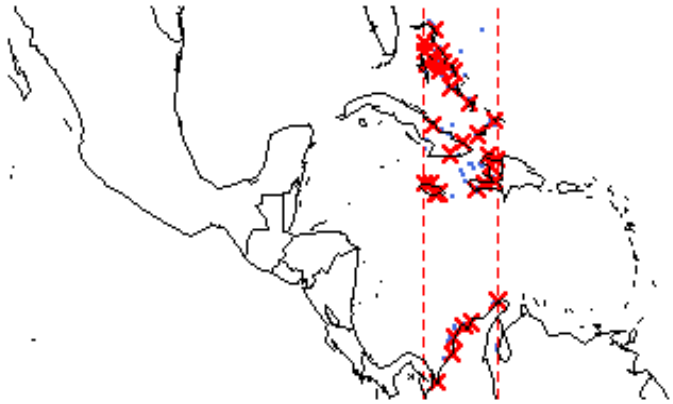

Figure 27. Location of Cluster Centers in Zone 18

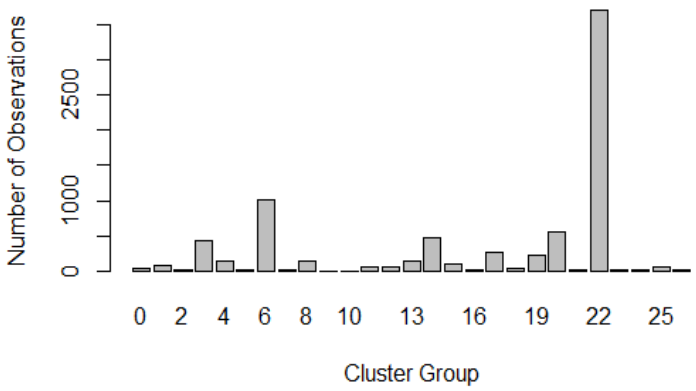

Figure 28. Distribution of Points by Cluster Group in Zone 19

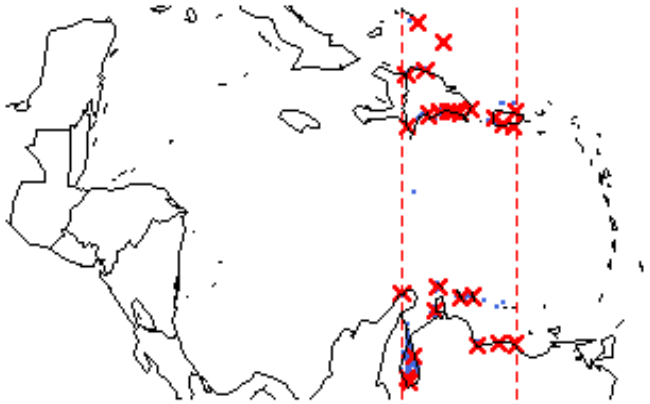

Figure 29. Location of Cluster Centers in Zone 19

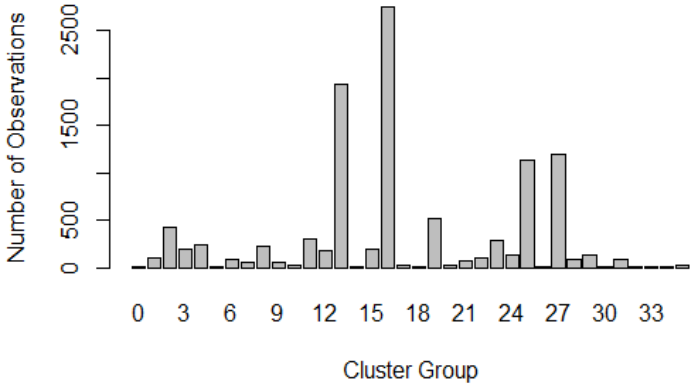

Figure 30. Distribution of Points by Cluster Group in Zone 20

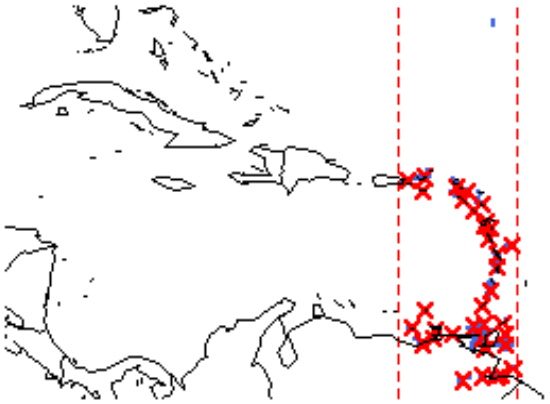

Figure 31. Location of Cluster Centers in Zone 20

In the set of Figures 18−31, there are cases where cluster centers are located over land. This is not an issue, as the cause of this occurrence stems from stop points surrounding a land mass. In Figures 18−31, it is possible to identify stop-points that are far from the cluster centers. These are the outlier, or cluster group 0, points that might warrant further investigation. For example the area near the border of Texas and Mexico in Figure 18 clearly shows a group of these outlier points. We also note zone 15, in Figure 21, has one very large cluster with the cluster center close to Port Fourchon, as does zone 16 in Figure 23. High traffic volume regions such as this can be extracted and further clustered. The main takeaway from these clustering results is that we are able to efficiently produce results that appear to be sound after visual inspection. We perform the clustering in a manner intended to save a great deal of computational time, and be easily replicated.

C. CLASSIFICATION TOOL RESULTS

We begin this process by splitting the dataset for each zone into a training and test set using an 80% and 20% split with random sampling, respectively. We then create a classification tree for each zone using the R package `rpart` (Therneau et al. 2015). The cluster group membership serves as the response variable for fitting each classification tree, and northing and easting serve as the explanatory variables. We prune the trees using cross-validation and the one standard error rule of Breiman et al. (1984). Once

pruned, the tree is used to predict the cluster group membership for the test sets. These values are then compared to the actual cluster group membership in order to find the overall misclassification rate for each UTM zone. A list of the complexity parameters we use to prune the final classification tree, as well as the final misclassification rate for each zone are shown in Table 6.

Table 6. Classification Tool Results

Zone	Complexity Parameter	Misclassification Rate
14	0.04000	0.01814
15	0.00771	0.00448
16	0.05730	0.01814
17	0.03800	0.08155
18	1e-5	0.04941
19	0.01620	0.03869
20	0.00620	0.03116

Table 6 shows that in all zones the misclassification rate is less than 9%. The very small misclassification rate in zone 15 can be attributed to a single, large cluster containing more than 90% of the stop-points in that zone. These results suggest that the classification tree, while having varying accuracy among zones, does demonstrate its usefulness. As we increase the number of observations for training the classification tree, the accuracy, as measured by the misclassification rate for each of the zones, should also increase.

THIS PAGE INTENTIONALLY LEFT BLANK

V. CONCLUSION

Due to the ever-increasing flow of AIS data into globally accessible databases, the amount of research being conducted in this field will continue to grow at a steady rate. While most anomaly detection and path prediction algorithms have had success in the past with dynamic data, there has not been much interest in analyzing the stop points for multiple vessels. We provide guidance on how to clean AIS data and define stop-points. Clustering stop points allows for a sizeable reduction in any dataset. By converting the stop points to northing and easting pairs and clustering by zone with OPTICS, this approach proves to be an efficient, timely technique to categorize a massive amount of dynamic data. The advantages of using OPTICS for clustering stop-points are its ability to identify clusters of different shapes and densities, and its ability to identify outliers that do not belong to any cluster. Furthermore, it is possible to construct a classification tool using the full dataset's clustering results and classification trees to identify stopping regions for new stop-points.

There are many concepts related to this topic that could be considered areas for future research. Although the treeClust method was unsuccessful in providing accurate clustering results, the treeClust method does show promise. Because treeClust inter-point distances are "learned" for a particular dataset, increasing the number of observations by using more than a month's worth of AIS data may give treeClust inter-point distances that yield better stop-point clusters. In addition, treeClust provides the means for combining other variables, including categorical ones such as information from the static AIS data, with geospatial locations. Another area of future research could be the construction of a fully automated system utilizing the techniques developed through this study. The automated system would use AIS data collected daily and compare predicted results from the classification tool compared to actual clustering results. In this case if there ever were to be a discrepancy between the two results, the vessel could be sorted for future investigation. This would provide a real-time solution to any vessels attempting to act in a nefarious manner.

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Ankerst M, Breunig M, Kriegel H, Sander J (1999) OPTICS: Ordering points to identify the clustering structure. *Proc. ACM SIGMOD'99 Int. Conf. on Management of Data* (Institute for Computer Science, Munich, Germany), 49-60.

Bay S (2017) Evaluation of factors on the patterns of ship movement and predictability of future ship location in the Gulf of Mexico. Master's thesis, Naval Postgraduate School, Monterey, CA.

Breiman L, Friedman J, Olshen R, Stone C (1984) *Classification and Regression Trees* (CRC Press, Wadsworth, NY).

Buttrey S (2016) treeClust: Cluster distances through trees. R package version 1.1-6. https://CRAN.R-project.org/package=treeClust

Buttrey S, Whitaker L (2016) treeClust: An R package for tree-based clustering dissimilarities. *The R Journal*. 7(2): 227–236.

Crewson P (2012) *Applied statistics handbook* (AcaStat Software, Winter Garden, FL).

de Selding P (2015) Harris, exactEarth to place AIS gear on iridium craft. *SpaceNews* (June 9), http://spacenews.com/harris-exactearth-to-place-ais-gear-on-iridium-craft/

Department of Homeland Security (2017a) Navigation Center: AIS requirements. Retrieved 19 July, https://www.navcen.uscg.gov/?pageName=AISRequirementsRev

Department of Homeland Security (2017b) Navigation Center: Types of AIS. Retrieved 19 July, https://www.navcen.uscg.gov/?pageName=typesAIS

Department of Homeland Security (2017c) Navigation Center: How AIS works. Retrieved 19 July, https://www.navcen.uscg.gov/?pageName=AISworks

Department of Homeland Security (2017d) Navigation Center: AIS messages. Retrieved 4 August, https://www.navcen.uscg.gov/?pageName=AISMessages

Ester M, Kriegel HP, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd*. 96(34):226–231.

Ginesi A (2009) ESA satellite receiver brings worldwide sea traffic tracking within reach. *ESA* (23 April), http://www.esa.int/Our_Activities/Space_Engineering_Technology/ESA_satellite_receiver_brings_worldwide_sea_traffic_tracking_within_reach

Google Earth (2015) Gulf of Mexico, 18° 27’ 57.14”N, 82° 03’ 48.56”W, Eye alt 2165 feet. *SIO, NOAA, U.S. Navy, NGA, GEBCO*. Retrieved 1 July, http://www.earth.google.com

Hahsler M, Piekenbrock M (2017) dbscan: Density based clustering of applications with noise (DBSCAN) and related algorithms. R package version 1.1-1. https://CRAN.R-project.org/package=dbscan

Hargreaves S (2010) Gulf oil spill: What’s at stake. *CNN Money* (30 May), http://money.cnn.com/2010/05/26/news/economy/gulf_economy/index.htm

International Maritime Organization (2017) AIS transponders. Retrieved 19 July, http://www.imo.org/en/OurWork/safety/navigation/pages/ais.aspx

Karney CF (2013) Algorithms for geodesics. *Journal of Geodesy*. 87(1):43–55.

Karney CF (2011) Transverse Mercator with an accuracy of a few nanometers. *Journal of Geodesy*. 85(8):475–485.

Kiremire AR (2011) The application of the Pareto Principle in software engineering. (Louisiana Tech University, Ruston).

MacQueen J (1967) Some methods for classification and analysis of multivariate observations. Le Cam LM, Neyman J, eds. *Proc. of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. (University of California, Berkeley, CA), 281–297.

Mao S, Tu E, Zhang G, Rachmawati L, Rajabally E, Huang G (2016) An automatic identification system (AIS) database for maritime trajectory prediction and data mining. Cao J, Cambria E, Lendasse A, Miche Y, Vong CM, eds. *Proceedings of ELM-2016*. (International Conference on Extreme Learning Machine 2016, Singapore), 241–257.

MarineTraffic (2017) Live map. Retrieved 21 July, http://www.marinetraffic.com/en/ais/home/centerx:-90.0/centery:17.1/zoom:4

National Research Council (1991) *Fishing Vessel Safety: Blueprint for a National Program* (The National Academies Press, Washington, DC).

Pallotta B, Vespe M, Bryan K (2013) Vessel pattern knowledge discovery from AIS data: A framework for anomaly detection and route prediction. *Entropy* 6(48): 2218–2245.

R Core Team (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/.

RStudio Team (2016) RStudio: Integrated Development for R. RStudio, Inc., Boston. http://www.rstudio.com/.

Schwehr K (2017) C++ decoder for Automatic Identification System for tracking ships and decoding maritime information. Retrieved 21 July, https://github.com/schwehr/libais

Shaham Y (2015) Visualizing mixed variable-type multidimensional data using tree distances. Master's thesis, Naval Postgraduate School, Monterey, CA.

Strauch K (2009) Atlantis leaves Columbus with a radio eye on Earth's sea traffic. *ESA* (4 December), http://www.esa.int/Our_Activities/Operations/i_Atlantis_i_leaves_Columbus_with_a_radio_eye_on_Earth_s_sea_traffic.

Therneau T, Atkinson B, Ripley B (2015) rpart: Recursive partitioning and regression trees. R package version 4.1-10. https://CRAN.R-project.org/package=rpart

THIS PAGE INTENTIONALLY LEFT BLANK

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California